DOCUMENT RESUME

ED 400 276 TM 025 528

AUTHOR Schnipke, Deborah L.
TITLE How Contaminated by Guessing Are Item-Parameter Estimates and What Can Be Done about It?
PUB DATE Apr 96
NOTE 19p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level; *Estimation (Mathematics); *Guessing (Tests); *Item Response Theory; Multiple Choice Tests; Simulation; *Test Items; *Timed Tests
IDENTIFIERS Item Characteristic Function; Item Parameters

ABSTRACT

When running out of time on a multiple-choice test, some examinees are likely to respond rapidly to the remaining unanswered items in an attempt to get some items right by chance. Because these responses will tend to be incorrect, the presence of "rapid-guessing behavior" could cause these items to appear to be more difficult than they really are. This study used simulated data from a normal distribution with a mean of 0 and a standard deviation of 1 for 5,000 examinees. It found that item response theory parameters are affected by rapid-guessing behavior, and that the Item Characteristic Curves (ICCs) were generally lower than the true ICCs. Using response times, an attempt was made to remove responses that appeared to be the result of rapid-guessing behavior. Two methods of removing responses were used. After removing the fast responses (rapid guesses), the item parameters and ICCs were recovered more accurately. The two methods of classifying responses as rapid guesses and removing them worked equally well in recovering the true parameters and ICCs. (Contains two tables, eight figures, and seven references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 400 276

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
DEBORAH SCHNIPKE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

How Contaminated by Guessing are Item-Parameter Estimates and What Can Be Done About It?

Deborah L. Schnipke
Law School Admission Council

BEST COPY AVAILABLE

A paper presented at the annual meeting of the
National Council on Measurement in Education
April, 1996, New York, NY

TM025528

How Contaminated by Guessing are Item-Parameter Estimates and What Can Be Done About It?

Abstract: When running out of time on a multiple-choice test, some examinees are likely to respond rapidly to the remaining unanswered items in an attempt to get some items right by chance. Because these responses will tend to be incorrect, the presence of "rapid-guessing behavior" could cause these items to appear more difficult than they really are. The present study used simulated data and found that item response theory parameters are affected by rapid-guessing behavior, and the item characteristic curves (ICCs) were generally lower than the true ICCs. Using response times, an attempt was made to remove responses that appeared to be the result of rapid-guessing behavior. Two methods of removing responses were used. After removing the fast responses (rapid guesses), the item parameters and ICCs were recovered more accurately. The two methods of classifying responses as rapid guesses and removing them worked equally well in recovering the true parameters and ICCs.

Both classical test theory and item response theory (IRT) assume that examinees answer items after fully considering them. An incorrect answer is taken to mean that the examinee was unable to answer the item (i.e., the item was too difficult for the examinee). On "speeded" (or "partially speeded") tests, performance may decline because the examinees run out of time. Examinees may begin to respond randomly, or after only briefly skimming the items. On such items, an incorrect answer does not necessarily mean that the item was too difficult for the examinee; the examinee may have been fully capable of answering the item correctly, given more time. Thus, the dimension of speed (rate of work) will affect scores on speeded tests.

Additionally, item parameters will be affected by random responding. If examinees respond randomly to items, these items will appear more difficult than they really are (e.g., Oshima, 1994). One reaction to this problem for test construction and other purposes is to use a different set of item parameters for items when they are administered at the end of a separately timed test section (because these items are typically the most affected by speededness). Another reaction is to hold item position constant when an item is reused (or from pretest, when initial item parameters are typically obtained, to the operational use of the items, when the items contribute to examinee scores).

These options may provide workable solutions in paper-and-pencil test construction and item analysis, but as testing programs consider moving to a computer-administered format, better solutions to the problem of random responding may exist. Better solutions will be especially important if the test will be given adaptively. In a computer adaptive test (CAT), an item can potentially be used in any position in the test. For a CAT to be successful, item parameters must not change depending on item position.

Although random responding is probably not the only cause of item-parameter instability, controlling for or eliminating random responding would probably lead to more stable and accurate parameter estimates. Yamamoto (1995) developed a model that takes

random responding into account when estimating item parameters. His model (HYBRID) assumes that examinees start a test by engaging in a strategy of thoughtful response, but at each successive item, some examinees switch to a strategy of random response. In his model, thoughtful responses are modeled by IRT. Under the random response strategy, the probability of a correct response is independent of examinee ability. HYBRID does not allow for more than one switch in strategy (as might happen if an examinee responds randomly only to a certain content area or item type), and on difficult items HYBRID may not be able to distinguish random responses from thoughtful responses if items are arranged in order of difficulty. However, using such a model is undoubtedly better than ignoring the random responses completely.

If one assumes that random responses are made quickly (perhaps as time expires), response times (if available) provide an additional way to identify responses as random. Using response times, Schnipke (1995) and Schnipke and Scrams (1996) showed that "rapid guesses" were obviously present on the analytical section of a computer-administered version of the GRE. The last half of the items showed evidence of a second underlying distribution of very fast, primarily incorrect, responses. Schnipke and Scrams (1996) modeled the response times with a two-state mixture model and showed that the response time distribution for each item could be described very well by a two-state model. The two states were labeled "rapid-guessing behavior" and "solution behavior," and the accuracy rates were consistent with the labels: the rapid-guessing state had a relatively flat, low accuracy rate that was near chance, and the solution state had a relatively flat, higher accuracy rate (determined by item difficulty, by definition); the accuracy rate in the area of overlap between the states was an increasing function of response time. The modeling techniques used by Schnipke and Scrams will be used in the present study, as described below.

In the present study, simulated data were used to address the questions "How contaminated by guessing are item-parameter estimates?" and "What can be done about it?" To address the first question (how contaminated), data were generated with and without rapid guesses, and the IRT parameters and item characteristic curves (ICCs) for the two conditions were compared to each other and to the true values. To address the second question (what can be done), two methods were used to remove what appeared to be rapid guesses (using the modeling techniques of Schnipke and Scrams, 1996), and the item parameters and ICCs were recomputed and were compared to the true values.

Method

Simulated examinees

Ability (θ) parameters were randomly sampled from a normal distribution with a mean of 0 and a standard deviation of 1 for 5000 simulated examinees.

Item parameters

Two types of parameters were simulated for 30 items: IRT parameters to describe the probability that a randomly chosen examinee of a given ability will answer the item correctly when engaged in solution behavior, and parameters to describe the response time distribution for each item.

IRT parameters

The unidimensional, three-parameter logistic (3PL) IRT model was used to describe the probability that a randomly chosen examinee of a given ability will answer the item correctly when engaged in solution behavior.[1] The 3PL model is given by

$$P_i(correct|\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}} \quad (1)$$

where θ is the examinee ability parameter,

$P_i(correct|\theta)$ is the probability that a randomly chosen examinee with ability θ answers item i correctly,

a_i is the discrimination parameter,

b_i is the difficulty parameter, and

c_i is the lower asymptote parameter (the probability that an examinee of very low ability will answer the item correctly).

The discrimination (a) parameters were randomly sampled from a normal distribution with a mean of 0.8 and a standard deviation of 0.3. The difficulty (b) parameters were randomly sampled from a normal distribution with a mean of 0 and a standard deviation of 1. The lower asymptote (c) parameters were randomly sampled from a uniform distribution ranging from .15 to .25 (roughly comparable to a 5-option multiple-choice item). The a, b, and c parameters were generated independently. The "items" were then sorted by the difficulty (b) parameter, from lowest to highest, to imitate a multiple-choice test with increasing item difficulty. The sorted items were then labeled 1 through 30, so that item 1 was the easiest and item 30 was the most difficult.

Response time parameters

Response time distributions for each item were based on work by Schnipke and Scrams (1996). Schnipke and Scrams found that a two-state mixture model (e.g., Luce, 1986) described the response time distribution for items on a speeded test very well. The two-state mixture model is given by

$$F_{Oi} = \rho_i F_{Gi} + (1 - \rho_i) F_{Si} \quad (2)$$

where F_{Oi} is the observed response-time distribution for item i,

ρ_i is the proportion of rapid guesses on item i,

F_{Gi} is the rapid-guessing response time distribution for item i, and

F_{Si} is the solution-behavior response time distribution for item i.

---

[1] Accuracy for rapid-guessing behavior was at chance (0.2), as discussed below, thus there are no IRT parameters are not necessary for the rapid-guessing state.

Schnipke and Scrams found that the rapid-guessing and solution behavior distributions could be described by lognormal distributions. The lognormal distribution is given by

$$F(t) = \frac{1}{\sqrt{t\sigma(2\pi)}} \times \exp\left\{\frac{-[\ln(t/m)]^2}{2\sigma^2}\right\} \quad (3)$$

where t is the response time (RT)

m is the scale parameter (the median of the RTs), and

σ is the shape parameter (the standard deviation of the ln(RT)'s; Evans, Hastings, & Peacock, 1993).

The present study also used a mixture of lognormal distributions. Thus, to specify the response time distribution for each item, 5 parameters are required: ρ, m_G and σ_G (for the rapid-guessing lognormal distribution), and m_S and σ_S (for the solution-behavior lognormal distribution). These 5 parameters were simulated for each of the 30 items (described next) and were used to generate the response times for each simulated examinee on each item (described below).

Proportion of rapid guesses (ρ). The proportion of rapid guesses for each item, ρ_i, was specified such that the proportions were very similar to what Schnipke and Scrams found in empirical data. The true proportion of rapid guesses and the recovered (estimated) proportion (to be discussed below) for each item are shown in Figure 1. The true proportion of rapid guesses was specified as an increasing function of item number (as was item difficulty, which created a positive relationship between item difficulty and the amount of rapid-guessing behavior).

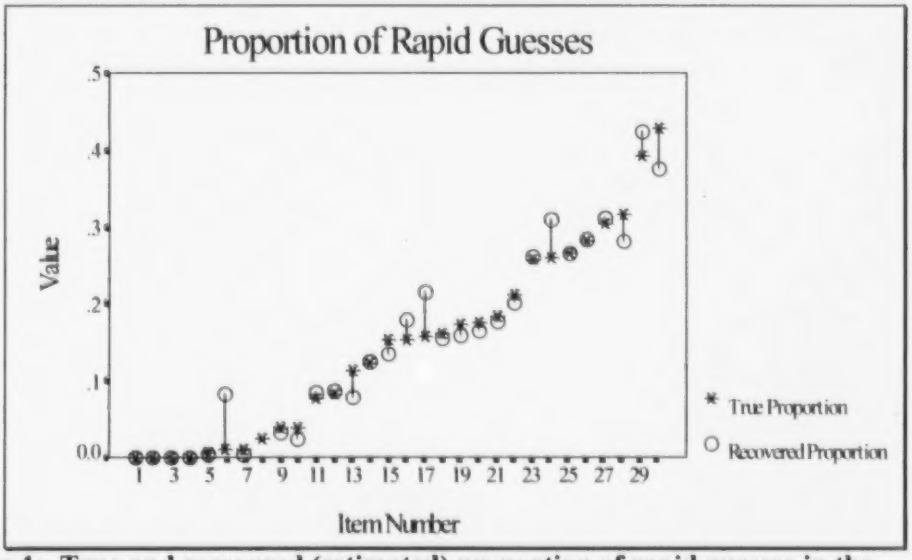

Figure 1: True and recovered (estimated) proportion of rapid guesses in the simulated data.

Rapid-guessing behavior distribution: m_G and σ_G. Schnipke and Scrams (1996) found that the rapid-guessing distribution could be constrained to be the same across all items and still fit the response times. They concluded that rapid-guessing behavior is essentially the same across items (i.e., it is independent of item characteristics). Thus in the present study, a common rapid-guessing distribution was used for all items with $m_G = 7.389$ sec, and $\sigma_G = 1.0$ ln(sec). (These values are similar to what Schnipke and Scrams found in real data.)

Solution behavior distribution: m_S and σ_S. The median and SD of the solution-behavior response time distribution varied across items, but were independent of item characteristics (such as difficulty, discrimination, lower asymptote, or item position). The median for each item, m_S, (which is expressed in sec) was sampled from a lognormal distribution with median of 60.34 sec and SD of 1.0 ln(sec), and σ_S (which is expressed in ln(sec)) was sampled from a lognormal distribution with median of 1.65 sec and SD of 0.08 ln(sec). A correlation of -0.3 between m_S and σ_S was built into the parameter selection as suggested by the results of Schnipke and Scrams (1996). The resulting solution behavior parameters are similar to those found by Schnipke and Scrams. The true and recovered (estimated) m_S and σ_S for each item are shown in Figure 2. (The recovered values will be discussed below.)

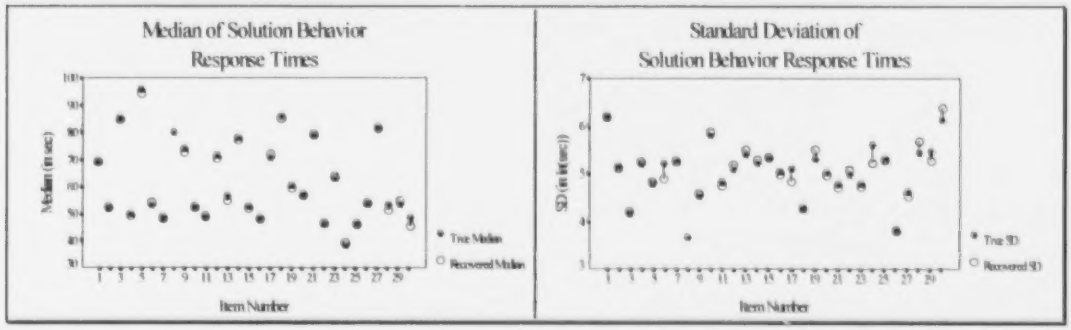

Figure 2: True and recovered (estimated) median and SD of the solution behavior response time distribution for each item.

Generating responses (accuracy) and response times

To generate responses (correct/incorrect) and response times for each examinee-item pair, it was first determined if the examinee was in the rapid-guessing or solution behavior state on the item, then accuracy and response times were generated, as described below. To determine in which state the examinee was, a random uniform number from 0 to 1 was generated. If this number was less than ρ (the proportion of rapid guesses on the item), the examinee was assigned to the rapid-guessing state for that item. Otherwise, the examinee was assigned to the solution state for the item. In this way, the proportion of examinees

assigned to the rapid-guessing state was approximately ρ. The assignment to the rapid-guessing or solution behavior state was independent of ability.

If the examinee was assigned to the solution behavior state, accuracy on the item was determined from a probabilistic application of the 3PL IRT model (Equation 1) based on the examinee's simulated ability (θ), and response time was randomly sampled from a lognormal distribution with the solution-behavior parameters for that item (shown in Figures 1 and 2 as the true values).

If the examinee was assigned to the rapid-guessing state, the probability of a correct response on the item was set equal to chance (0.2), and response time was randomly sampled from a lognormal distribution with the rapid-guessing parameters which were the same for all items, as discussed above. A dataset of 0's and 1's (incorrect/correct) and response times was thus created.[2]

Classifying responses as rapid guesses

To address the second question (what can be done about item-parameter contamination), two methods were used to classify the fast responses as rapid guesses (ignoring, of course, the true classifications used to generate the data). The responses that were classified as rapid guesses were then removed during item parameter estimation to see if the true item parameters could be recovered more accurately.

To classify responses as rapid guesses, two techniques were used: one that was probabilistic, and one that was based on a cutoff. Both techniques relied on fitting a two-state mixture model to the response times for each item. Thus before discussing the classification techniques, the modeling of the response times will be discussed.

Modeling the response times

Nonlinear regression (SPSS, 1994) was used to fit the two-state mixture model (Equation 2) to the response time distribution for each item. The underlying distributions (the rapid-guessing and solution behavior distributions) were specified as lognormal distributions (Equation 3). The rapid-guessing distribution was not constrained to be the same across items during the estimation of the two-state model parameters (although the data were generated that way). Thus 5 parameters were free to vary for each item: $\hat{\rho}$ (the estimated proportion of rapid guesses), $\hat{m}_G$ and $\hat{\sigma}_G$ (the lognormal parameter estimates for the rapid-guessing distribution), and $\hat{m}_S$ and $\hat{\sigma}_S$ (the lognormal parameter estimates for the solution-behavior distribution).

---

[2] A dataset was also created with ρ = 0 for all items (i.e., no rapid guesses). All examinees were in the solution behavior state on all items; accuracy followed the 3PL model and response times followed the solution behavior distribution (with m_S and σ_S).

The true and recovered (estimated) parameters are shown in Figures 1-3. As shown in Figure 1, the proportion of rapid guesses was recovered fairly well, although the estimated proportion was sometimes a little high and sometimes a little low. As shown in Figure 2, the median, m_S, of the solution behavior distribution for each item was recovered quite well, although σ_S was recovered a little less well for some items. The items on which σ_S was recovered least well were generally the same items on which the proportion of rapid guesses was recovered least well (items 6, 13, 17, 24, 28, 29, and 30).

As shown in Figure 3, the median, m_G, of the rapid-guessing behavior distribution for most items was recovered well. The exceptions are item 6, where $\hat{m}_G$ was far too high (which is why the estimated proportion of rapid guesses on item 6 was too high), and some of the items toward the beginning of the test. Likewise, σ_G for items toward the beginning of the test was not recovered well, although σ_G was recovered well for items toward the end of the test. This may be explained by the fact that the true proportion of rapid guesses was very small at the beginning of the test so the rapid-guessing distribution could not be estimated well, whereas toward the end of the test the proportion of rapid guesses was large enough to provide stable estimates of the rapid-guessing distribution.

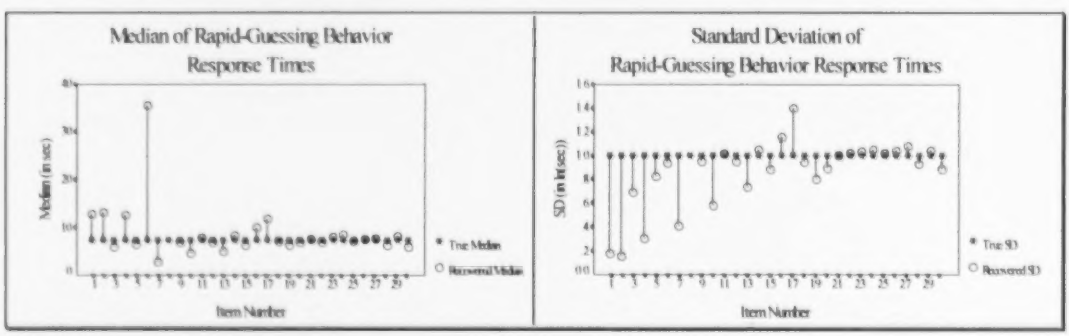

Figure 3: True and estimated median and SD of the rapid-guessing behavior distribution for each item.

Classification methods

The estimated parameters (shown as the recovered parameters in Figures 1-3) were used to classify responses as rapid guesses using the following two methods.

Responses classified probabilistically. The first method of classifying responses as rapid guesses attempted to emulate the response time distribution that would have arisen if there had been no rapid-guessing behavior. That is, the method attempted to sample a solution behavior distribution from the mixture distribution. Based on the mixture model, the proportion of responses at a given response time that should be in each distribution was calculated. For the responses at each response time, the responses were randomly assigned to the two distributions to match the expected proportions. In this sense, the method was probabilistic; there was some probability that a given response would be assigned to a particular state (rapid-guessing or solution behavior). The responses that were actually rapid guesses and the responses that were classified as rapid guesses were not exactly the same, of course, but the proportion of responses assigned to each distribution were correct according to the model. (The random assignment to the two underlying distributions was independent of accuracy, so although the proportion of responses classified followed the model, the accuracy rates were not quite right.) The proportion of responses classified as rapid guesses is shown in Table 1 (as is the actual proportion[1] for comparison).

Responses classified based on a cutoff. The second method of classifying responses as rapid guesses was based on a cutoff. The cutoffs were established based on the two-state mixture model. The recovered parameters were used to determine, for each item, where the two underlying distributions crossed, weighted by the proportion of examinees in each distribution (ρ_i or $1-\rho_i$). The point at which distributions cross is where a response is equally likely to come from either the rapid-guessing or solution behavior distribution. This point was used as the cutoff. Once the cutoff for each item was established, all responses with a response time less than the cutoff were classified as rapid guesses, and all responses with a response time greater than the cutoff were classified as solution behavior responses. The cutoffs that were used are given in Table 1, as are the proportions of responses that were classified as rapid guesses using this method.

[1] The actual proportion of rapid guesses in the data set is not ρ (although they are very similar). ρ was used to generate the proportion of rapid guesses, but because of random variability, the numbers are not exactly the same.

Table 1: Actual proportion and classified proportions of rapid guesses in the simulated data, and the cutoff (in sec) used for the cutoff-based approach.

Item	Actual Proportion	Proportion Classified as Rapid Guesses		Value of Cutoff (in sec)
		Probabilistic Method	Cutoff Method	
1	0	0	0	1
2	0	0	0	1
3	0	0	0	1
4	.1	0	0	1
5	.7	.6	.6	16.7
6	1.2	7.7	1.7	14.8
7	.9	.5	.5	6.6
8	2.6	0	0	0
9	3.6	3.2	2.9	18.1
10	3.8	2.4	2.4	9.3
11	7.8	8.5	6.8	14.5
12	9.0	8.6	7.7	18.0
13	10.8	8.3	7.8	12.7
14	12.1	12.5	10.7	21.1
15	15.4	13.4	12.0	14.7
16	14.9	17.7	12.7	16.0
17	15.8	22.0	15.8	24.6
18	15.5	15.6	14.6	27.5
19	17.9	15.8	15.0	16.4
20	17.5	16.6	15.3	17.6
21	17.8	18.0	16.3	25.0
22	20.8	19.8	16.7	15.3
23	25.5	26.2	23.2	23.0
24	27.2	29.9	24.3	15.2
25	26.8	26.4	22.5	16.0
26	28.1	28.2	25.6	23.0
27	30.1	31.4	28.8	30.0
28	31.5	28.2	25.8	16.6
29	39.3	42.7	37.7	22.3
30	42.7	38.1	35.5	15.6

Treatment of the responses classified as rapid guesses

The responses that were classified as rapid guesses were treated as if the item that gave rise to the so-called rapid guess had never been presented to that examinee. To do this, the response code was changed to the "never presented" code (which is not technically true, but it produces the desired result). In this way, responses that appear to be rapid guesses do not influence parameter estimation. Responses that were classified as solution behavior were not altered. The recoding of responses classified as rapid guesses was done independently (not additively) for the two methods of classification, of course.

Parameter estimation

BILOG (Mislevy & Bock, 1990) was used to estimate item parameters for the 4 conditions:

- no rapid guessing (no rapid guessing was simulated during data generation),
- including rapid guessing (rapid-guessing behavior was simulated during data generation),
- rapid-guessing removed probabilistically (rapid-guessing behavior was simulated during data generation; some responses were classified as rapid guesses using the probabilistic method and were removed), and
- rapid-guessing removed with a cutoff (rapid-guessing behavior was simulated during data generation; some responses were classified as rapid guesses based on a cutoff and were removed).

The third and fourth conditions (where rapid-guessing was removed) were modifications of the second condition (which included rapid guessing). That is, the exact same responses were used for each simulated examinee-item pair, except for the responses that were classified as rapid guesses. For responses classified as rapid guesses, the response code was changed to indicate that the examinee had not received that item, as discussed above.

In order to set the scale so that the item parameter estimates from BILOG could be compared, the final estimated ability distribution was scaled to the standard normal. For the condition that includes rapid guessing, forcing the ability distribution to be standard normal will not reflect the true distribution. Although the original ability distribution was standard normal, the inclusion of rapid-guessing behavior distorts the original distribution so that it is no longer standard normal. Scaling the final estimated ability distribution to the standard normal in this case is not "correct" in the sense that we know the original was distorted. However, in real test data, we would not necessarily know that speed (and hence rapid-guessing behavior) was a factor influencing test scores, and we would scale the final estimated ability distribution to the standard normal. Thus, for the simulated data which included rapid guesses, the ability estimates were scaled as would be done with real test data, and these results were compared to the case in which there was no rapid guessing and to the cases in which an attempt was made to remove rapid guesses.

The IRT item parameters were estimated for each item under each of the 4 conditions. It was expected that the "no rapid guessing" condition would recover the true item parameters very well, that the "including rapid guessing" condition would not recover the item parameters as well at the end of the test where the proportion of rapid guessing is the highest. How much of a difference there is between these conditions answers the first research question ("How contaminated by guessing are item-parameter estimates?")

The second research question (what can be done about the contamination?) is addressed by the last two conditions where some responses were classified as rapid guesses and were then removed. The point of this was to remove responses that were likely to be rapid guesses based on response times (ignoring the true classifications which were used to generate that data and which would not be known in real test data) to see if the true item parameters could be recovered more accurately.

Results

Figure 4 shows the true difficulty (b) parameters, as well as the 4 estimated b parameters for each item (1 estimated value for each of the 4 conditions). When rapid guesses were included in the parameter estimation, the items toward the end of the test appeared more difficult (higher $\hat{b}$ value) than they really are. When there were no rapid guesses and when responses classified as rapid guesses were removed (using either method), the true difficulty parameters were recovered quite well, as shown in Figure 4. As a test-level index of how well the difficulty parameters were recovered, the correlation between the true and estimated b's was calculated. As shown in Table 2, the correlation between b and $\hat{b}$ was lowest when rapid guesses were included, highest when there were no rapid guesses, and in between when rapid guesses were removed. All of the correlations were quite high, though, for the difficulty (b) parameter.

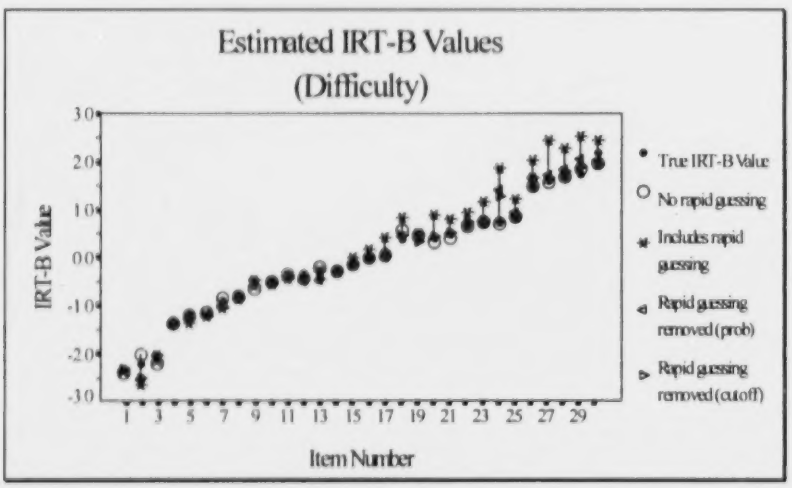

Figure 4: True and estimated difficulty parameter for each item.

Table 2: Correlations between the true and estimated IRT parameters for the four conditions.

	No Rapid Guessing	Including Rapid Guessing	Rapid Guessing Removed	
			Probabilistic Method	Cutoff Method
Difficulty (*b*)	.9970	.9888	.9922	.9938
Discrimination (*a*)	.9841	.7876	.9208	.9305
Lower asymptote (*c*)	.5398	.3502*	.5692	.6036

*Not significantly different from 0 at α=.05.

Figure 5 shows the true and estimated discrimination (*a*) parameters for each item. The true discrimination parameters generally were not recovered as well as the difficulty parameters, as can be seen in Table 2 (the correlations are lower than those for the *b* parameter). When no rapid guesses were present, the discrimination parameters were recovered fairly well. When rapid guesses were included, the discrimination parameters were recovered the least well (as shown by the correlation between a and $\hat{a}$ and by Figure 5). The discrimination parameter was underestimated on most items when rapid guesses were present. When the responses classified as rapid guesses were removed (using either method), the discrimination parameters were recovered better than when rapid guesses were included, but not as well as the "no rapid guessing" condition.

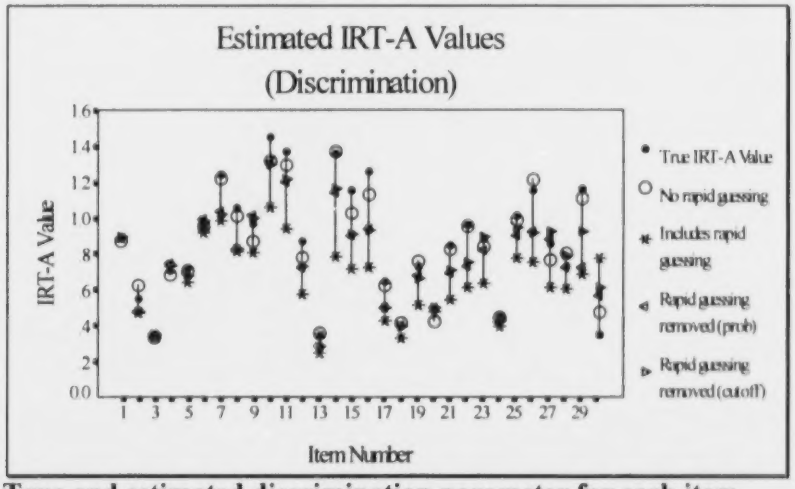

Figure 5: True and estimated discrimination parameter for each item.

Figure 6 shows the true and estimated lower asymptote (c) parameters for each item. The true lower asymptote parameters were not recovered as well as the difficulty or discrimination parameters (e.g., the correlations between c and $\hat{c}$, shown in Table 2, are lower than those for difficulty or discrimination). As with the difficulty and discrimination parameters, however, when rapid guesses were included, the lower asymptote parameters were recovered the least well. When the responses classified as rapid guesses were removed (using either method), the lower asymptote parameters were recovered at least as well (if not better) as when no rapid guesses were present, as shown in Figure 6 and Table 2.

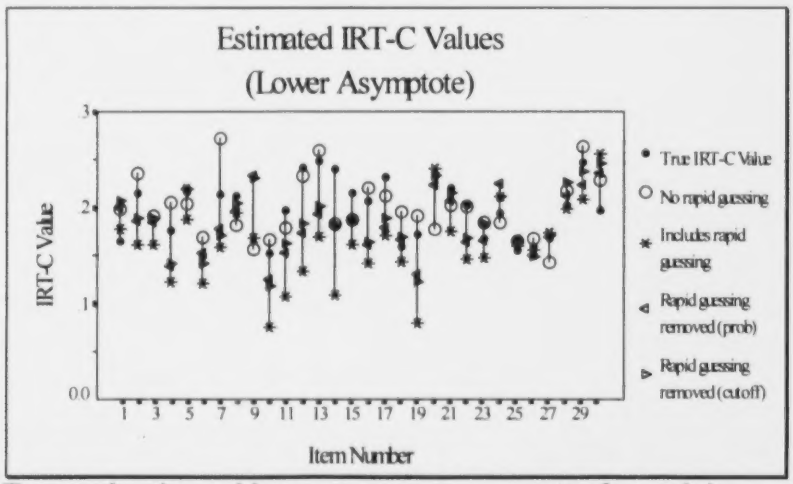

Figure 6: True and estimated lower asymptote parameter for each item.

To compare the combined effects of the IRT a, b, and c parameters, the item characteristic curve (ICC) for each item was calculated. The ICC provides the probability that an examinee with a given ability will answer the item correctly. Figure 7 shows the true and estimated ICCs for several items throughout the simulated test (items 10, 15, 20, 25, and 30). Figure 8 shows the residuals of the estimated ICCs (compared to the true ICC) for the same 6 items. As shown in Figures 7 and 8, when rapid guesses were included, the ICC was artificially low for items toward the end of the test; an examinee appears to have a smaller probability of answering an item correctly than is really the case. This, of course, is analogous to the difficulty parameter being too high. (The slopes of the ICCs were also altered by rapid-guessing behavior, which is related to the a's and c's being recovered less well.) When there were no rapid guesses, or when responses that were classified as rapid guesses were removed (using either method), the ICCs were recovered very well.

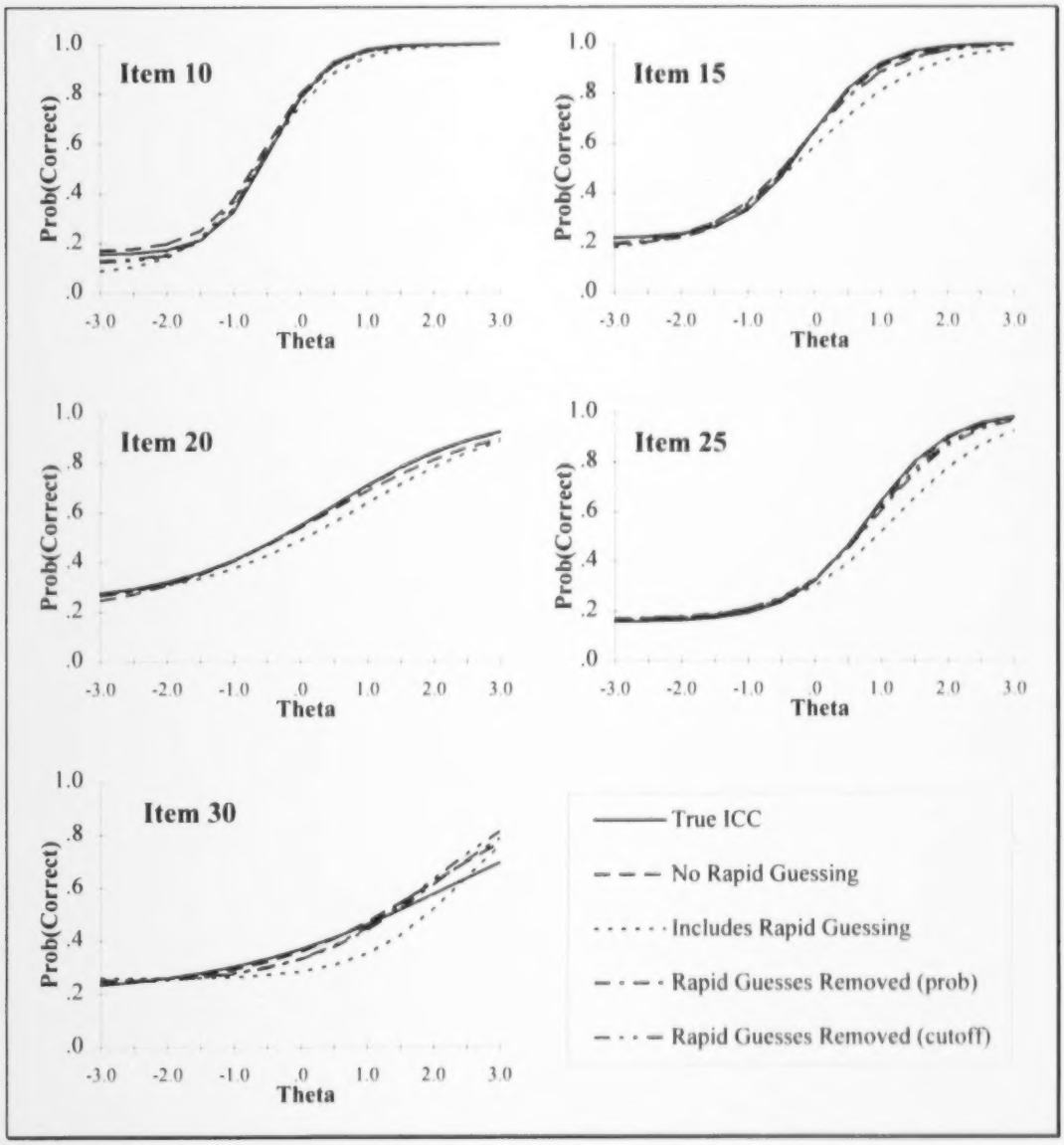

Figure 7: Estimated and true item characteristic curves for several simulated items.

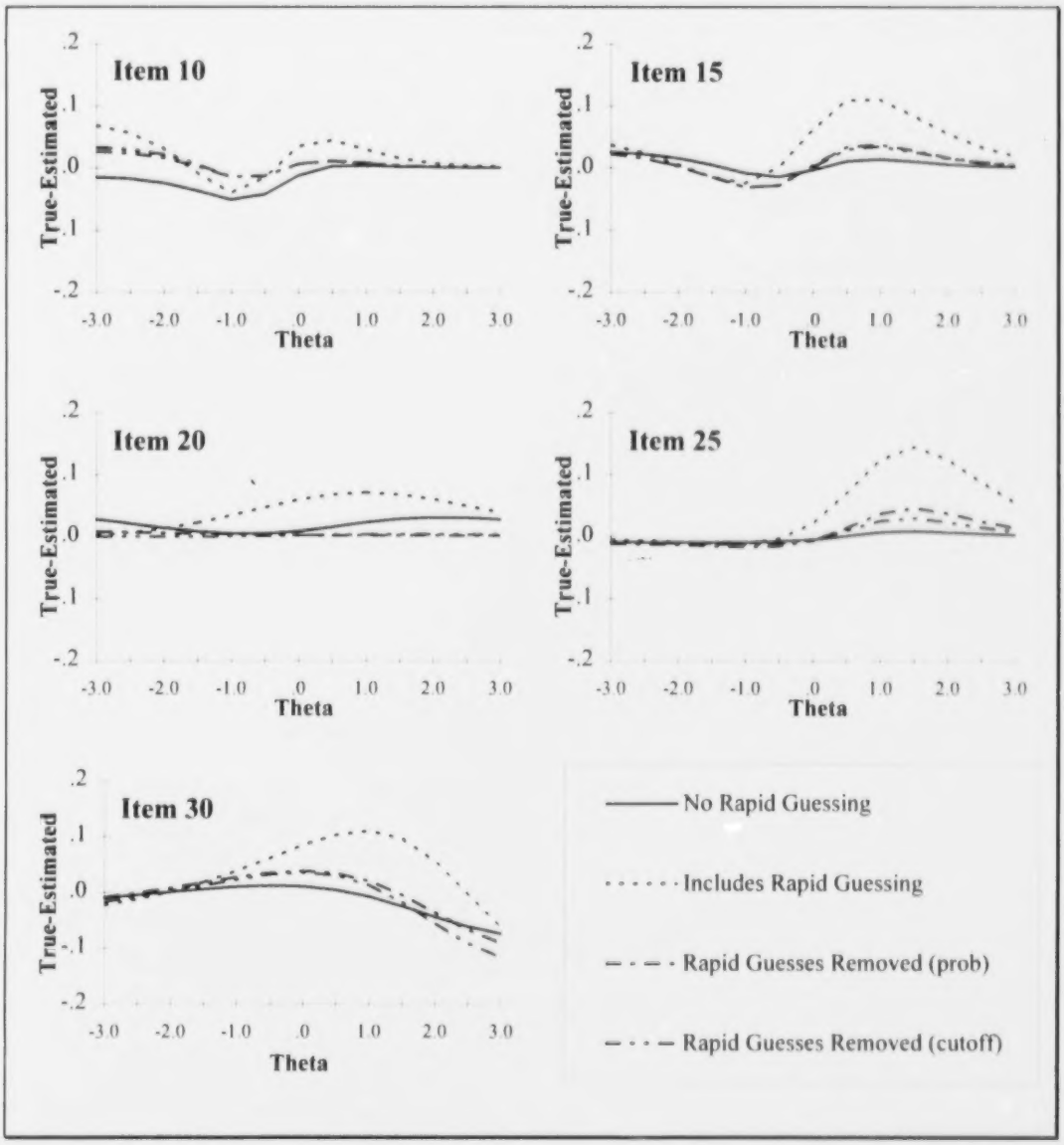

Figure 8: Residuals of the estimated item characteristic curves for several items.

Discussion

On speeded tests, some examinees are likely to respond very quickly to items as time expires in the hopes of getting some items right by chance. This is often a wise strategy for examinees to use, but these rapid responses cause problems for practitioners who analyze test data. These fast responses, or rapid guesses as I have been calling them, will contaminate item-parameter estimates if they are included during parameter estimation. As shown in the present study, when rapid guesses are present, items appear more difficult and less discriminating than they really are.

When tests are administered on computers, response times can be collected. Response times provide clues about which responses are rapid guesses and which are not. In the present study, response times were simulated to match distributions found by Schnipke and Scrams (1996) and were used to classify responses as rapid guesses to see if removing these responses would provide more accurate parameter estimates. A simulation was needed to do this because the true parameters needed to be known.

Two methods were used to classify responses as rapid guesses. One attempted to emulate a solution-behavior-only distribution by sampling from the mixture distribution. Some, but not all, of the fastest response times were removed so that the remaining number of responses at each response time approximately matched the number expected by the underlying solution behavior distribution. The other method used a cutoff to classify responses; responses made faster than the cutoff were classified as rapid guesses. This method truncated the distribution, whereas the first method did not.

The two methods for classifying responses as rapid guesses produced virtually identical results. The parameter estimates were recovered equally well with the two methods. Because the cutoff approach is easier to implement and easier to understand, the cutoff approach is recommended, although the other approach works well, too.

When new tests are constructed, the new tests may not have the desired psychometric properties if inaccurate item parameters (e.g., ones derived at the end of a speeded test) are used. This may be a problem especially in adaptive tests because item selection and examinee ability estimation depend heavily on item parameters. If response times are available, it is recommended that practitioners look for evidence of rapid-guessing behavior in their data and correct for it if it is there, especially if the item parameters will be used in adaptive tests.

There are several limitations of the present study; these include (1) ability and rapid-guessing behavior were generated independently, (2) solution-behavior response time distributions were generated independently of item difficulty, and (3) the proportion of rapid guesses and item difficulty both increased as a function of item position. For generalizing the results, the biggest limitation is probably that ability and rapid-guessing behavior were simulated independently. Ability and rapid-guessing behavior probably are related in real data, but because the exact nature of this relationship is not known, no dependency was incorporated.

Oshima (1994) found that random guessing distorts difficulty estimates the least when items are arranged in order of ascending difficulty (if random guessing also

increases toward the end of the test). This was the case in the present study. The difficulty estimates would be expected to be more distorted by rapid guesses if items were not increasingly difficult toward the end of the test, and in this situation it might be even more important to counteract the effects of rapid-guessing behavior by attempting to removed such responses.

References

Evans, M., Hastings, N., & Peacock, B. (1993). *Statistical distributions*. New York: John Wiley & Sons.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models [Computer software and manual]. Chicago: Scientific Software, Inc.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*(3), 200-219.

Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

Schnipke, D. L., & Scrams, D. J. (1996, June). *Modeling response time in testing with a two-state mixture model: A new approach to detect speededness*. Paper to be presented at the meeting of the Psychometric Society, Banff, Alberta, Canada.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model*. (TOEFL Technical Report No. TR-10). Princeton, NJ: Educational Testing Service.